	QMRF identifier (JRC Inventory): **To be entered by JRC**
	QMRF Title: **Algae Chronic (NOEC) Toxicity model (IRFMN)-v. 1.0.1**
	Printing Date: October **2022**

1.QSAR identifier

1.1. QSAR identifier (title):

Algae Chronic (NOEC) Toxicity model (IRFMN)-v. 1.0.1

1.2. Other related models:

Alga acute, daphnia acute daphnia chronic fish acute fish chronic

1.2. Software coding the model:

VEGA (https://www.vegahub.eu/)

The VEGA software provides QSAR models to predict tox, ecotox, environ, phys-chem and toxicokinetic properties of chemical substances.

emilio.benfenati@marionegri.it

2.General information

2.1. Date of QMRF:

October 2022

2.2. QMRF author(s) and contact details:

Emilio Benfenati Istituto di Ricerche Farmacologiche Mario Negri - IRCSS Via Mario Negri 2, 20156 Milano, Italy emilio.benfenati@marionegri.it https://www.marionegri.it/

2.3. Date of QMRF update(s):

NA

2.4. QMRF update(s):

NA

2.5. Model developer(s) and contact details:

Alberto Manganaro Istituto di Ricerche Farmacologiche Mario Negri -IRCSS Via Mario Negri 2, 20156 Milano, Italy alberto.manganaro@marionegri.it https://www.marionegri.it/

2.6. Date of model development and/or publication:

NA

2.7. Reference(s) to main scientific papers and/or software package:

[1] Benfenati E, Manganaro A, Gini G

VEGA-QSAR: AI inside a platform for predictive toxicology

Proceedings of the workshop "Popularize Artificial Intelligence 2013", December 5th 2013, Turin, Italy

Published on CEUR Workshop Proceedings Vol-1107

[2] Benfenati E., Lombardo A. (2020) VEGAHUB for Ecotoxicological QSAR Modeling. In: Ecotoxicological QSARs, Part of the Methods in Pharmacology and Toxicology book series (MIPT), Springer Protocols, Editor Kunal Roy, pages 759-787.

[3] Benfenati E, Roncaglioni A, Lombardo A, Manganaro A. Integrating QSAR, Read-Across and Screening Tools: The VEGAHUB Platform as an Example. In: Advances in Computational Toxicology: Methodologies and Applications in Regulatory Science, Hong H, Ed. Springer Nature, 2019. pp 365-382.

2.8. Availability of information about the model:

The model is non-proprietary and the training set is available.

2.9. Availability of another QMRF for exactly the same model:

Another QMRF is not available.

3.Defining the endpoint - OECD Principle 1

3.1. Species:

The species name is Rhapidocelis subcapitata, previously named as Selenastrum capricornutum or Pseudokirchneriella subcapitata

3.2. Endpoint:

ECOTOX 6.1.5. Long-term toxicity to aquatic algae and cyanobacteria. OECD, Test No. 201: Freshwater Alga and Cyanobacteria, Growth Inhibition Test.

3.3. Comment on endpoint:

NA

3.4. Endpoint units:

mg/L

3.5. Dependent variable:

Original data were transformed from mg/l to mmol/l using a box-cox transformation. Data falling outside the range (mean of the box-cox transformed values) ± 3*(standard deviation) were excluded.

3.6. Experimental protocol:

NOEC 72h (growth rate), OECD TG 201

3.7. Endpoint data quality and variability:

410 experimental data on algae chronic toxicity (NOEC, 72h growth rate) retrieved from the Japanese Ministry of Environment (http://www.env.go.jp/en/chemi/sesaku/aquatic_Mar_2016.pdf) and selected ac-cording to the OECD TG 201 requirement. The dataset was split into training (328 substances) and test set (82 subtances)

4.Defining the algorithm - OECD Principle 2

4.1. Type of model:

The Alga Chronic (NOEC) toxicity model (IRFMN) –v.1.0.1 is based on 410 experimental data on algae chronic toxicity (NOEC, 72h growth rate) retrieved from the Japanese Ministry of Environment (http://www.env.go.jp/en/chemi/sesaku/aquatic_Mar_2016.pdf) and selected ac-cording to the OECD TG 201 requirement. The model is a Tree Ensemble Random Forest.

4.2. Explicit algorithm:

Tree Ensemble Random Forest

To derive the models, we divided the data in training and test sets with the ratio of 80:20. In order to obtain a uniform distribution of the endpoint values between the two subsets we applied an activity and descriptors sampling method. We performed a Principal Component Analysis (PCA) on all the descriptors and we selected the first two principal components. We selected five random compounds, and then we picked the most dissimilar compound from the sample pool according to the first two principal components and the response using several combinations of distance metrics and scoring functions. Then we added the compound to the pool repeating the operation until we reached the desired number for the training set.

Among the several algorithms used, we obtained the best results in terms of performance with a Random Forest called Tree ensemble. Tree ensemble builds a series of regression trees with different rows and different variables (according to certain parameters) and then it aggregates the results as an ensemble of models. It chooses the parameters for the variables of each tree and the number of compounds evaluating the performance of several models (Hyperparameter tuning Research) using as metric R^2 of a Bootstrap (100 iterations) cross-validation on training set.

4.3. Descriptors in the model:

The model uses descriptors calculated with DRAGON, and these descriptors are then used by the Random Forest algorithm.

4.4. Descriptor selection:

In order to select the variables, we used two methods implemented in R packages for each dataset: the genetic algorithm (gaselect package) and the Variable Selection Using Random Forest (VSURF) package. We imported both the pools of variables of each dataset into a KNIME workflow to derive the models

4.5. Algorithm and descriptor generation:

The descriptors are those obtained originally by DRAGON, and the selected ones were implemented in VEGA

4.6. Software name and version for descriptor generation:

DRAGON Calculation of several sets of molecular descriptors from molecular geometries (topological, geometrical, WHIM, 3D-MoRSE, molecular profiles, etc.) http://www.disat.unimib.it/chm

Prof. R.Todeschini -distributed by Talete srl, via Pisani 13, 20124 Milano, Italy

4.7. Chemicals/Descriptors ratio:

Each tree in the Random Forest applies a much smaller set of descriptors

5.Defining the applicability domain - OECD Principle 3

5.1. Description of the applicability domain of the model:

The Applicability Domain (AD) is assessed using the original algorithm implemented within VEGA. An overall AD index is calculated, based on a number of parameters, which relate to the results obtained on similar chemicals within the training and test sets and is defined in this way for this QSAR model´s predictions:

If 1 ≥ AD index > 0.85, the predicted substance is regarded in the Applicability Domain of the model

If 0.85 ≥ AD index > 0.7, the predicted substance could be out of the Applicability Domain of the model

If AD index ≤ 0.7, the predicted substance is regarded out of the Applicability Domain of the model
Indices are calculated on the first k = 2 most similar molecules, each having S_k similarity value with the target molecule.

5.2. Method used to assess the applicability domain:

The Applicability Domain and the chemical similarity is measured with the algorithm developed for VEGA. Full details are in the VEGA website (www.vegahub.eu), including the open access paper describing it [6]. The VEGA AD also evaluates the correctness of the prediction on similar compounds (accuracy), the consistency between the predicted value for the target compound and the experimental values of the similar compounds, the range of the descriptors, and the presence of unusual fragments, using atom centered fragments.

Similar molecules with known experimental value:

This index takes into account how similar are the first two most similar compounds found. Values near 1 mean that the predicted compound is well represented in the dataset used to build the model, otherwise the prediction could be an extrapolation. Defined intervals are:

If 1 ≥ index > 0.85, strongly similar compounds with known experimental value in the training set have been found.

If 0.85 ≥ index > 0.7, only moderately similar compounds with known experimental value in the training set have been found.

If index ≤ 0.7, no similar compounds with known experimental value in the training set have been found.

Accuracy (average error) of prediction for similar molecules:

This index takes into account the classification accuracy in prediction for the two most similar compounds found. Values near 1 mean that the predicted compounds fall in an area of the model's space where the model gives reliable predictions (no misclassifications), otherwise the lower is the value, the worse the model behaves. Defined intervals are:

If index < 0.8, accuracy of prediction for similar molecules found in the training set is good

If 1.5 > index ≥ 0.8, accuracy of prediction for similar molecules found in the training set is not optimal

If index ≥ 1.5, accuracy of prediction for similar molecules found in the training set is not adequate

Concordance for similar molecules:

This index takes into account the difference between the predicted value and the experimental values of the two most similar compounds. Values near 0 mean that the prediction made disagrees with the values found in the model's space, thus the prediction could be unreliable. Defined intervals are:

If index < 0.8, molecules found in the training set have experimental values that agree with the target compound predicted value

If 1.5 > index ≥ 0.8, similar molecules found in the training set have experimental values that slightly disagree with the target compound predicted value

If index ≥ 1.5, similar molecules found in the training set have experimental values that completely disagree with the target compound predicted value

Maximum error of prediction among similar molecules:

This index takes into account the maximum error in prediction among the two most similar compounds. Values near 0 means that the predicted compounds fall in an area of the model's space where the model gives reliable predictions without any outlier value. Defined intervals are:

If index < 0.8, the maximum error in prediction of similar molecules found in the training set has a low value, considering the experimental variability

If 1.5 > index ≥ 0.8, the maximum error in prediction of similar molecules found in the training set has a moderate value, considering the experimental variability

If index ≥ 1.5, the maximum error in prediction of similar molecules found in the training set has a high value, considering the experimental variability

Atom Centered Fragments similarity check: This index takes into account the presence of one or more fragments that aren't found in the training set, or that are rare fragments. First order atom centered fragments from all molecules in the training set are calculated, then compared with the first order atom centered fragments from the predicted compound; then the index is calculated as following: a first index RARE takes into account rare fragments (those who occur less than three times in the training set), having value of 1 if no such fragments are found, 0.85 if up to 2 fragments are found, 0.7 if more than 2 fragments are found; a second index NOTFOUND takes into account not found fragments, having value of 1 if no such fragments are found, 0.6 if a fragments is found, 0.4 if more than 1 fragment is found. Then, the final index is given as the product RARE * NOTFOUND. Defined intervals are:

If index = 1, all atom centered fragment of the compound have been found in the compounds of the training set

If 1 > index ≥ 0.7, some atom centered fragment of the compound have not been found in the compounds of the training set or are rare fragments

If index < 0.7, a prominent number of atom centered fragments of the compound have not been found in the compounds of the training set or are rare fragments

Model descriptors range check:

This index checks if the descriptors calculated for the predicted compound are inside the range of descriptors of the training and test set. The index has value 1 if all descriptors are inside the range, 0 if at least one descriptor is out of the range. Defined intervals are:

Index = TRUE, descriptors for this compound have values inside the descriptor range of the compounds of the training set

Index = FALSE, descriptors for this compound have values outside the descriptor range of the compounds of the training set

5.3. Software name and version for applicability domain assessment:

VEGA (www.vegahub.eu)

5.4. Limits of applicability:

VEGA provides a quantitative value for the prediction of each substance. This helps the user to identify potential critical aspects, which are indicated. Similar compounds are shown.

6.Internal validation - OECD Principle 4

6.1. Availability of the training set:

Yes

6.2. Available information for the training set:

CAS RN: Yes

Chemical Name: No

Smiles: Yes

Formula: No

INChI: No

MOL file: No

NanoMaterial: No

6.3. Data for each descriptor variable for the training set:

No

6.4. Data for the dependent variable for the training set:

NA

6.5. Other information about the training set:

For each dataset, to further refine the results, we applied a pruning process both to the compounds and to the descriptors pools. Firstly, we removed the compounds for which it was not feasible to calculate AlogP (Ghose-Crippen octanol-water partition coefficient (Ghose and Crippen, 1986; Viswanadhan et al., 1993; Ghose et al., 1998)), as it is generally well acknowledged that this descriptor is the most correlated to the response. Then, to reduce the great number of variables, we removed all the descriptors with constant values (var(X) =0), or which correlate over 0.95 (Pearson) with at least one another descriptor. To derive the models, we divided the data in training and test sets with the ratio of 80:20. In order to obtain a uniform distribution of the endpoint values between the two subsets we applied an activity and descriptors sampling method. We performed a Principal Component Analysis (PCA) on all the descriptors and we selected the first two principal components. We selected five random compounds, and then we picked the most dissimilar compound from the sample pool according to the first two principal components and the response using several combinations of distance metrics and scoring functions. Then we added the compound to the pool repeating the operation until we reached the desired number for the training set.

Among the several algorithms used, we obtained the best results in terms of performance with a Random Forest called Tree ensemble. Tree ensemble builds a series of regression trees with different rows and different variables (according to certain parameters) and then it aggregates the results as an ensemble of models. It chooses the parameters for the variables of each tree and the number of compounds evaluating the performance of several models (Hyperparameter tuning Research) using as metric R2 of a Bootstrap (100 iterations) cross-validation on training set

6.6. Pre-processing of data before modelling:

Data curation

SMILES:

Firstly, we generated the SMILES structures from the chemical name and CAS RN for each substance using ChemCell (2019) and Marvin View (Marvin 17.28.0, 2012017, ChemAxon, 2019). We manually checked the correspondence and correctness among the obtained structures, chemical name and CAS RN among several websites and public database like ChemIDplus Advanced (NIH, 2019), PubChem (NCBI, 2019), ChemSpider (Royal Society of Chemistry, 2019), DSSTox. Then, we added several structures, which have not automatically generated.

We normalized the SMILES with istMolBase 1.0.3. (in-house software), then we neutralized them using KNIME 3.5. Since pH is a critical issue in the experimental assays on algae, we considered ionized normalized SMILES and we calculated the major microspecies at pH 7.5 and 8.1 using JChem for Excel. We removed the compounds for which the SMILES changed depending on pH (in range 7.5-8.1).

We cleaned the datasets excluding the following compounds: metal complexes, inorganics, mixtures of structural isomers, ambiguous structures, non-ionic surfactant mixtures, complex disconnected structures (e.g. polymers), chemicals whose correspondence name-CAS was not found, UVCB, salts; only the acid form was kept.

Values cleaning:

We selected continuous experimental values excluding those reported as a range, greater/less than a certain threshold, or approximate values. We converted each experimental value from mg/l to mmol/l, on the basis of the molecular weight calculated from the chemical structure. We also removed the compounds for which the experimental toxicity values were higher than the experimental water solubility values. For this purpose we retrieved the experimental water solubility values mainly from a large database of more than 4,000 chemicals that we pruned in the LIFE project ANTARES and from GuideChem and Sigma-Aldrich websites in the case we didnot find the water solubility values elsewhere.

Dealing with multiple values:

To deal with multiple continuous data we referred to the procedures described in ECHA guidance R.10 (2008) for ecotoxicological continuous endpoints. In case the experimental conditions and the reliability of the studies were the same, we considered the ratio between the maximum and the minimum values; if it was higher than one log unit we eliminated the data. Then, we calculated the median, the arithmetic and geometric mean in mmol/l to check if there were differences among them. We found a very good correlation (R2 close to 1) between the values of each combination (arithmetic vs geometric mean, arithmetic mean vs median, geometric mean vs median) and finally the geometric mean was preferred (ECHA guidance R.10, 2008).

To normalize the data we performed two types of transformation, the logarithm of the geometric mean and the Box-cox transformation. Since the box-cox transformation gave better results in terms of normalization of the data, it was finally used to normalize the data. We excluded data falling outside the range (mean of the box-cox transformed values) ± 3*(standard deviation). Concerning the final chronic fish toxicity, once we merged all the dataset, the final dataset increased up to 94 chemicals

6.7. Statistics for goodness-of-fit:

Training set: n 328, RMSE 0.79, R2 0.91

6.8. Robustness - Statistics obtained by leave-one-out cross-validation:

NA

6.9. Robustness - Statistics obtained by leave-many-out cross-validation:

NA

6.10. Robustness - Statistics obtained by Y-scrambling:

NA

6.11. Robustness - Statistics obtained by bootstrap:

NA

6.12. Robustness - Statistics obtained by other methods:

NA

7.External validation - OECD Principle 4

7.1. Availability of the external validation set:

Yes

7.2. Available information for the external validation set:

CAS RN: Yes

Chemical Name: No

Smiles: Yes

Formula: No

INChI: No

MOL file: No

NanoMaterial: No

7.3. Data for each descriptor variable for the external validation set:

NA

7.4. Data for the dependent variable for the external validation set:

NA

7.5. Other information about the external validation set:

NA

7.6. Experimental design of test set:

Test set: n 82, RMSE 1.79, R2 0.51

Test set in AD: n 13, RMSE 0.74, R2 0.59

Test set could be out of AD: n 31, RMSE 1.35, R2 0.68

Test set out of AD: n 38, RMSE 2.29, R2 0.37

7.7. Predictivity - Statistics obtained by external validation:

NA

7.8. Predictivity - Assessment of the external validation set:

NA

7.9. Comments on the external validation of the model:

NA

8.Providing a mechanistic interpretation - OECD Principle 5

8.1. Mechanistic basis of the model:

NA

8.2.A priori or a posteriori mechanistic interpretation:

A posteriori only

8.3. Other information about the mechanistic interpretation:

NA

9.Miscellaneous information

9.1. Comments:

NA

9.2. Bibliography:

[1]Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32. https://link.springer.com/article/10.1023/A:1010933404324

[2]Genuer, R., Poggi, J. M., &Tuleau-Malot, C. (2018). VSURF: Variable Selection Using Random Forests. R package version 1.0.4. URL https://CRAN.R-project.org/package=VSURF https://journal.r-project.org/archive/2015/RJ-2015-018/RJ-2015-018.pdf

[3]Benfenati E., Lombardo A. (2020) VEGAHUB for Ecotoxicological QSAR Modeling. In: Ecotoxicological QSARs, Part of the Methods in Pharmacology and Toxicology book series (MIPT), Springer Protocols, Editor Kunal Roy, pages 759-787.

[4]Benfenati E, Roncaglioni A, Lombardo A, Manganaro A. Integrating QSAR, Read-Across and Screening Tools: The VEGAHUB Platform as an Example. In:Advances in Computational Toxicology:Methodologies and Applications in Regulatory Science, Hong H, Ed. SpringerNature, 2019. pp 365-382

[5] OECD, Test No. 201: Freshwater Alga and Cyanobacteria, Growth Inhibition Test. Paris: Organisation for Economic Co-operation and Development, 2011. Accessed: Nov. 02, 2022. [Online]. Available: https://www.oecd-ilibrary.org/environment/test-no-201-alga-growth-inhibition-test_9789264069923-en

[6] Floris M, Manganaro A, Nicolotti O, Medda R, Mangiatordi GF, Benfenati E. A generalizable definition of chemical similarity for read-across. J Cheminform. 2014 Oct 18;6(1):39. doi: 10.1186/s13321-014-0039-1. PMID: 25383097; PMCID: PMC4212147

9.3. Supporting information:

Training set(s)Test set(s)Supporting information:

All available dataset are present in the model inside the VEGA software

10.Summary (JRC QSAR Model Database)

10.1. QMRF number:

To be entered by JRC

10.2. Publication date:

To be entered by JRC

10.3. Keywords:

To be entered by JRC

10.4. Comments:

To be entered by JRC